Lost in space? Generalising subtree prune and regraft to spaces of phylogenetic networks

Magnus Bordewich[a], Simone Linz[b], Charles Semple[c]

[a]*School of Engineering and Computing Sciences, Durham University, Durham DH1 3LE, United Kingdom.*
[b]*Department of Computer Science, University of Auckland, New Zealand.*
[c]*School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand.*

Abstract

Over the last fifteen years, phylogenetic networks have become a popular tool to analyse relationships between species whose past includes reticulation events such as hybridisation or horizontal gene transfer. However, the space of phylogenetic networks is significantly larger than that of phylogenetic trees, and how to analyse and search this enlarged space remains a poorly understood problem. Inspired by the widely-used rooted subtree prune and regraft (rSPR) operation on rooted phylogenetic trees, we propose a new operation—called subnet prune and regraft (SNPR)—that induces a metric on the space of all rooted phylogenetic networks on a fixed set of leaves. We show that the spaces of several popular classes of rooted phylogenetic networks (e.g. tree child, reticulation visible, and tree based) are connected under SNPR and that connectedness remains for the subclasses of these networks with a fixed number of reticulations. Lastly, we bound the distance between two rooted phylogenetic networks under the SNPR operation, show that it is computationally hard to compute this distance exactly, and analyse how the SNPR-distance between two such networks relates to the rSPR-distance between rooted phylogenetic trees that are embedded in these networks.

Keywords: phylogenetic networks, reticulation-visible network, subnet

Email addresses: m.j.r.bordewich@durham.ac.uk (Magnus Bordewich), s.linz@auckland.ac.nz (Simone Linz), charles.semple@canterbury.ac.nz (Charles Semple)

prune and regraft, tree-based network, tree-child network

1. Introduction

Searching through tree space is a key ingredient to many popular algorithms that reconstruct an optimal phylogenetic tree from a set of molecular sequence data. With the ever-increasing size of available data sets and given that the number of possible trees grows exponentially with the size of the leaf sets, analysing the mathematical properties of tree space continues to be an active area of research (e.g. Allen and Steel (2001); Bordewich and Semple (2004); Bryant (2004); Gordon et al. (2013); Owen and Provan (2011); Sanderson et al. (2011); Whidden and Matsen (2015)). This began at least in the early seventies when Robinson (1971) laid the foundation for the popular graph-theoretic nearest neighbour interchange (NNI) operation that induces a metric on the space of phylogenetic trees. Together with the local tree rearrangement operations of subtree prune and regraft as well as tree bisection and reconnection (Swafford et al., 1996), NNI was indispensable to the successful development of leading tree reconstruction methods (Bouckaert et al., 2014; Guindon et al., 2010; Ronquist and Huelsenbeck, 2003; Stamatakis, 2006).

In contrast, the space of phylogenetic networks remains poorly understood although it is now widely acknowledged that rooted leaf-labeled digraphs with underlying cycles are better suited to represent complex evolutionary histories that include reticulation events such as hybridisation and horizontal gene transfer (Gusfield, 2014; Huson et al., 2010). While several metrics on different classes of rooted phylogenetic networks have recently been developed (e.g. see Cardona et al. (2009a,b); Nakhleh (2010)), almost none of these metrics imposes a natural structure on the space of phylogenetic networks. It is precisely the structure on tree space and the efficient computation of a so-called tree neighbourhood that the above-mentioned tree reconstruction methods take advantage of in order to search tree space. To date, the only available metric that induces some structure on the space of phylogenetic networks and that can therefore be used to traverse phylogenetic network space in search of an optimal (rooted) phylogenetic network is rather ad hoc and based on a search that works in layers from trees to networks of increasing complexity (Yu et al., 2013, 2014). Indeed, the mathematical properties of this search (e.g. does it find each network within each layer) are

unknown. Additionally, for unrooted phylogenetic networks, two operations have recently been developed that generalise the NNI operation from trees to networks (Huber et al., 2016a,b). While the first operation (Huber et al., 2016a) induces a metric on a relatively simple class of phylogenetic networks that do not have any overlapping cycles, the operation presented in (Huber et al., 2016b) draws its inspiration from the cubic graph literature and can be used to transform any unrooted phylogenetic network into any other such network.

Given the lack of biologically-motivated and well-studied rearrangement operations on phylogenetic networks, it is consequently unsurprising that many algorithms that reconstruct a rooted phylogenetic network from sequence data under some optimisation criterion (e.g. parsimony (Fischer et al., 2015; Jin et al., 2009) or likelihood (Jin et al., 2006a)) refrain from searching phylogenetic network space. Instead, they often employ a two-step workaround (Jin et al., 2006a,b) that consists of reconstructing a phylogenetic tree and adding a fixed number of internal edges to the tree such that the resulting network is in some sense optimal.

This paper contributes to filling the lack of methods to reconstruct phylogenetic networks directly from molecular sequence data. In particular, we introduce a rearrangement operation—called subnet prune and regraft (SNPR)—on rooted phylogenetic networks that has its motivation in the popular rooted subtree prune and regraft operation on rooted phylogenetic trees (Bordewich and Semple, 2004). In comparison to the operations introduced in Huber et al. (2016a,b), a SNPR operation can move a subnetwork across a greater distance (i.e. an arbitrary number of edges) in a network. As we will see, generalising rearrangement operations from phylogenetic trees to networks opens up a set of interesting and novel questions for further investigations. For example, we not only show that SNPR induces a metric on the space of all rooted phylogenetic networks but also that several well-studied classes of rooted phylogenetic networks are connected under this operation, that is, starting at any network in the class, we can 'move' to any other network in the class by applying a sequence of SNPR operations such that the resulting network after each operation is also in the class.

The paper is organised as follows. The next section contains notation and terminology that is used throughout this paper. Section 3 introduces the SNPR operation and establishes several of its properties. The main result of this section is that SNPR induces a metric on the space of all rooted phylogenetic networks on a given leaf set. In Sections 4-6, we show that the

spaces of tree-child, tree-based, and reticulation-visible networks with a fixed number of reticulations are connected as well as the space of networks that embed a given set of rooted phylogenetic trees. For two given phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ on X, we then turn to computing their SNPR-distance, i.e. the minimum number of SNPR operations that transform $\mathcal{N}$ into $\mathcal{N}'$. In Section 7, we show that, in general, computing this distance is an NP-hard problem and bound the minimum number from above. We also analyse how the SNPR-distance between two rooted phylogenetic networks relates to the rSPR-distance between rooted phylogenetic trees that are embedded in these networks. Throughout the paper, we allow for a general rooted phylogenetic network to have edges in parallel. There are compelling reasons for this and we described them in Section 8, where we also present a modified SNPR operation that does not allow for parallel edges. We end the paper with some concluding remarks.

2. Preliminaries

This section provides notation and terminology that is used in the remainder of this paper. Throughout the paper, X always denotes a finite set. A (rooted) *phylogenetic network on* X is a rooted acyclic directed graph with the following properties:

(i) the unique root has out-degree two;

(ii) vertices with out-degree zero have in-degree one, and the set of vertices with out-degree zero is X; and

(iii) all other vertices have either in-degree one and out-degree two or in-degree two and out-degree one.

For technical reasons, we allow a single-root vertex to be a phylogenetic network. The observant reader will have noticed that we allow edges to be in *parallel*, i.e. we allow two edges to join the same pair of two distinct vertices. Allowing parallel edges is atypical. However, in the context of this paper, there are compelling reasons for this allowance as we explain in Section 8.

For a phylogenetic network $\mathcal{N}$ on X, the vertices of out-degree zero, that is the elements in X, are called *leaves* and X is the *leaf set*. Furthermore, the vertices of in-degree one and out-degree two are *tree vertices*, while the vertices of in-degree two and out-degree one are *reticulations*. An edge (u, v)

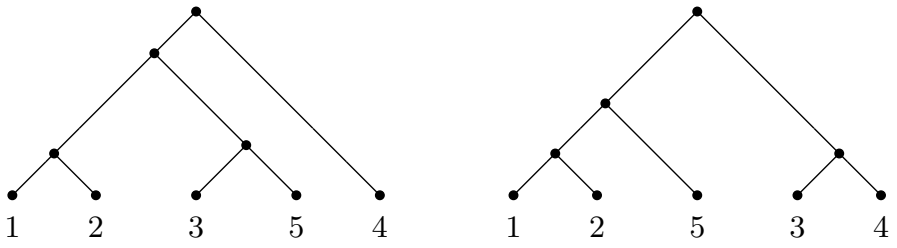

Figure 1: Two rooted binary phylogenetic trees on five leaves. Both trees are displayed by the phylogenetic network that is shown in the middle of Fig. 2.

in $\mathcal{N}$ is a *reticulation edge* if v is a reticulation; otherwise, (u, v) is a *tree edge*. Note that, for a reticulation v, we do not distinguish between the two reticulation edges directed into v. Moreover, for two vertices u and v in $\mathcal{N}$, we say that u is a *parent* of v and v is a *child* of u precisely if there is an edge (u, v) in $\mathcal{N}$.

A *rooted binary phylogenetic X-tree* is a phylogenetic network on X with no reticulations. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. For a subset X' of X, we denote by $\mathcal{T}|X'$ the rooted binary phylogenetic X'-tree obtained from the minimal rooted subtree of $\mathcal{T}$ that connects all leaves in X' by contracting non-root degree-two vertices.

Let $\mathcal{N}$ be a phylogenetic network on X. Let u and v be vertices in $\mathcal{N}$, and let e be an edge in $\mathcal{N}$. We say that v (respectively, e) is a *descendant* of u or, equivalently, *below* u if there is a directed path in $\mathcal{N}$ starting at u and traversing v (respectively, both end vertices of e). Also, a directed path in $\mathcal{N}$ starting at a vertex u and ending at a leaf ℓ is called a *tree path* if every vertex, other than u and ℓ, is a tree vertex. Now let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. We say that $\mathcal{N}$ *displays* $\mathcal{T}$ if, up to contracting non-root degree-two vertices, $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting edges and vertices, in which case, the resulting digraph (which has no underlying cycles) is an *embedding* of $\mathcal{T}$ in $\mathcal{N}$. To illustrate, Fig. 1 shows two rooted binary phylogenetic trees that are both displayed by the phylogenetic network that is depicted in the middle of Fig. 2.

Three particular classes of networks are becoming increasingly prominent in the literature. We describe these next. Let $\mathcal{N}$ be a phylogenetic network on X with root ρ. A vertex v in $\mathcal{N}$ is *visible* if there is a leaf ℓ in $\mathcal{N}$ such that every directed path from ρ to ℓ traverses v. Visibility is an attractive property as it allows the present to 'see' the past. For example, the three

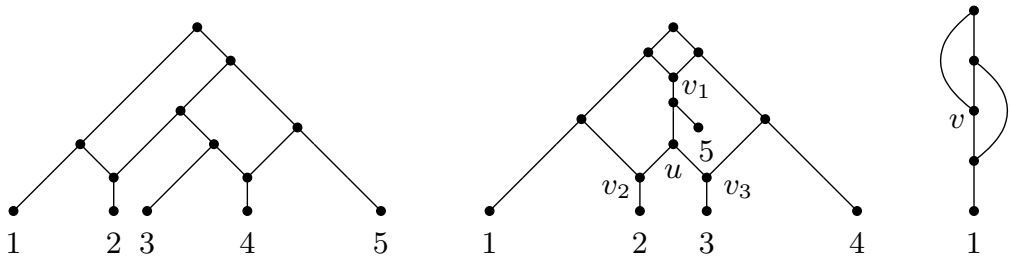

Figure 2: Left: A tree-child and, hence, reticulation-visible and tree-based network. Middle: A reticulation-visible and, hence, tree-based network that is not tree child since u has two reticulation children. Right: A tree-based network that is not reticulation visible since v is not visible and, hence, not tree child.

reticulations v_1, v_2, and v_3 of the network that is shown in the middle of Fig. 2 are visible since each directed path from the root to 5 traverses v_1, each directed path from the root to 2 traverses v_2 and, similarly for 3 and v_3. On the other hand, the reticulation v of the network that is shown in the right of the same figure is not visible because there exists a directed path from the root of the network to the only leaf 1 that does not traverse v. Now suppose that $\mathcal{N}$ has no edges in parallel. We say that $\mathcal{N}$ is *tree child* if every vertex in $\mathcal{N}$ is visible. Furthermore, $\mathcal{N}$ is *reticulation visible* if every reticulation in $\mathcal{N}$ is visible. Lastly, up to allowing parallel edges, $\mathcal{N}$ is *tree based* if there is an embedding $\mathcal{S}$ of a rooted binary phylogenetic X-tree $\mathcal{T}$ in $\mathcal{N}$ that covers every vertex in $\mathcal{N}$, in which case, $\mathcal{S}$ (as well as $\mathcal{T}$) is a *base tree* for $\mathcal{N}$. Note that the phylogenetic tree that is shown in the left of Fig. 1 is a base tree of the network that is shown in the middle of Fig. 2 while the tree that is shown in the right of Fig. 1 is not a base tree of that network. It immediately follows that every tree-child network is reticulation visible, and it can be shown that every reticulation-visible network is tree based (Francis and Steel, 2015). Among the aforementioned three classes of phylogenetic networks, only tree-based networks were introduced in a way that explicitly allows for parallel edges to be present (Francis and Steel, 2015). Examples of tree-child, reticulation-visible, and tree-based networks are shown in Fig. 2.

Finally, let $e = (u, v)$ be a reticulation edge in a phylogenetic network $\mathcal{N}$ on X, and suppose that u is not a reticulation. The *deletion* of e in $\mathcal{N}$, denoted $\mathcal{N}\backslash e$, is the operation of deleting e, and contracting u and v. Observe that $\mathcal{N}\backslash e$ is also a phylogenetic network on X.

Lemma 2.1. *Let $\mathcal{N}$ be a tree-child, reticulation-visible, or tree-based network, respectively, on X with at least one reticulation. Then there is a reticulation edge e in $\mathcal{N}$ such that $\mathcal{N}\backslash e$ is a tree-child, reticulation-visible, or tree-based network on X, respectively.*

Proof. We prove the lemma for when $\mathcal{N}$ is a reticulation-visible network on X. The proof for when $\mathcal{N}$ is tree child is simpler and omitted. Furthermore, the proof for when $\mathcal{N}$ is tree based follows from the fact that tree-based networks are allowed to have parallel edges (Francis and Steel, 2015).

Let v be a reticulation in $\mathcal{N}$ with the property that there is no reticulation in $\mathcal{N}$ that is an ancestor of v. As $\mathcal{N}$ is acyclic, such a choice is possible. Let u be a parent of v, and let e denote the edge (u, v). Suppose that $\mathcal{N}\backslash e$ has a pair of parallel edges. Then these edges arise as a result of the contraction of u or v. Let f and g denote the edges in $\mathcal{N}\backslash e$ resulting from the contraction of u and v, respectively. Since $\mathcal{N}$ has no parallel edges, f and g are distinct.

If g is in parallel to another edge in $\mathcal{N}\backslash e$, then the unique child of v in $\mathcal{N}$ is a reticulation; a contradiction, as otherwise, v is not visible in $\mathcal{N}$. So g is not in parallel with another edge in $\mathcal{N}\backslash e$. Therefore f is in parallel in $\mathcal{N}\backslash e$. This implies that the child vertex of u that is not v, say v', in $\mathcal{N}$ is also a reticulation. Furthermore, as f is in parallel in $\mathcal{N}\backslash e$, it follows that (t, v') is an edge in $\mathcal{N}$, where t is the unique parent of u. Moreover, as v has no ancestor that is a reticulation, v' has no such ancestor either. It is now easily seen that $\mathcal{N}\backslash(t, v')$ has no parallel edges and is reticulation visible. This completes the proof of the lemma. □

Suppose that $\mathcal{N}$ is a tree-child, reticulation-visible, or tree-based network. In view of Lemma 2.1, there is an ordering $(v_1, v_2, \ldots, v_k)$ of the reticulations of $\mathcal{N}$ such that, setting $\mathcal{N} = \mathcal{N}_0$, for all $i \in \{1, 2, \ldots, k\}$, if $\mathcal{N}_{i-1}$ is a tree-child, reticulation-visible, or tree-based network, then there is an edge, e_i say, directed into v_i with $\mathcal{N}_i = \mathcal{N}_{i-1}\backslash e_i$ being tree child, reticulation visible, or tree based, respectively. We call $(e_1, e_2, \ldots, e_k)$ a *deletion ordering* of $\mathcal{N}$.

3. Subnet Prune and Regraft

In this section, we introduce the subnet prune and regraft operation on phylogenetic networks and establish some of its basic properties. Let $\mathcal{N}$ be a phylogenetic network on X. For the convenience of defining this operation, view the root ρ of $\mathcal{N}$ as a vertex of in-degree zero and out-degree one adjacent

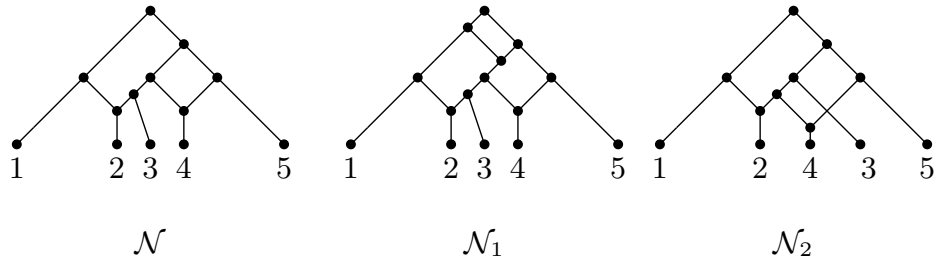

Figure 3: A phylogenetic network $\mathcal{N}$ and two phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ that can be obtained from $\mathcal{N}$ by applying (III) and (I), respectivley, to $\mathcal{N}$. Note that neither $\mathcal{N}_1$ nor $\mathcal{N}_2$ is tree child.

to the original root of $\mathcal{N}$, that is, view ρ as a *pendant root*. Let $e = (u, v)$ be an edge in $\mathcal{N}$, and consider the following three operations acting on $\mathcal{N}$:

(I) If u is a tree vertex (and not equal to ρ), then delete e, contract u, subdivide an edge that is not a descendant of v with a new vertex u', and add the new edge (u', v).

(II) If u is a tree vertex and v is a reticulation, then delete e, and contract u and v.

(III) Subdivide e with a new vertex v', subdivide an edge in the resulting network that is not a descendant of v' with a new vertex u', and add the new edge (u', v').

It is easily seen that applying any one of (I)–(III) to $\mathcal{N}$ results in a phylogenetic network on X. We say that a phylogenetic network on X has been obtained from $\mathcal{N}$ by a single *subnet prune and regraft* (SNPR) if it can be obtained by applying exactly one of (I)–(III) to $\mathcal{N}$. To illustrate, Fig. 3 shows two phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ that can be obtained from the phylogenetic network $\mathcal{N}$ that is shown in the same figure by applying (III) and (I), respectively. The well-known operation of rooted subtree prune and regraft (rSPR) is a certain application of SNPR. Specifically, this application is the restriction of $\mathcal{N}$ to a rooted binary phylogenetic X-tree and only allowing operation (I). Lastly, while applying (I) to a phylogenetic network $\mathcal{N}$ results in a phylogenetic network with the same number of reticulations, applying (II) (resp. (III)) to $\mathcal{N}$ results in a phylogenetic network whose number of

reticulations is decreased (resp. increased) by one.

Remarks.

1. The convenience of viewing ρ as a pendant root in the definition is so that we do not have to make a special case of the operation that undoes "pruning" an edge that is incident with the root.

2. As mentioned in the introduction, Yu et al. (2014) have recently described an operation to search phylogenetic network space in layers. Although the authors did not mention how to deal with parallel edges that may arise as a result of their operation, it is clear that their operation is similar to SNPR, which was developed independently and with the main motivation of generalising rSPR to phylogenetic networks. In fact, their operation can be regarded as a generalisation of an SNPR operation since they not only allow for the switching of a parent (as we do in (I)), but also for the switching of a child reticulation. Roughly speaking, such an operation replaces a reticulation edge (u, v) with a reticulation edge (u, w). As we will see in the following, the three possibilities in the definition of an SNPR operation are sufficient to establish that SNPR induces a metric on the space of phylogenetic networks as well as on several popular network classes.

It is natural to want each of the operations (I)–(III) to be *reversible*, that is, if a phylogenetic network $\mathcal{N}'$ on X can be obtained from $\mathcal{N}$ by a single SNPR, then $\mathcal{N}$ can be obtained from $\mathcal{N}'$ by a single SNPR. The following proposition shows that this is indeed the case.

Proposition 3.1. *Let $\mathcal{N}$ and $\mathcal{N}'$ be two phylogenetic networks on X. Suppose that $\mathcal{N}'$ is obtained from $\mathcal{N}$ by a single* SNPR*. Then, up to isomorphism, $\mathcal{N}$ can be obtained from $\mathcal{N}'$ by a single* SNPR*.*

Proof. If $\mathcal{N}'$ is obtained from $\mathcal{N}$ by applying one of (I), (II), and (III), then it is easily seen that, up to isomorphism, $\mathcal{N}$ can be obtained from $\mathcal{N}'$ by applying (I), (III), and (II), respectively. □

To illustrate Proposition 3.1, observe that the two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}_1$ that are shown in Figure 3 can be obtained from one another by applying a single SNPR and, likewise, the two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}_2$ that are shown in the same figure.

Now, let $\mathcal{N}$ and $\mathcal{N}'$ be phylogenetic networks on X. A SNPR-*sequence* for $\mathcal{N}$ and $\mathcal{N}'$ is a sequence

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_t = \mathcal{N}'$$

of phylogenetic networks on X such that, for all $i \in \{1, 2, \ldots, t\}$, we have $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by a single SNPR, in which case, we say that there is a SNPR-sequence *connecting* $\mathcal{N}$ and $\mathcal{N}'$. The *length* of the SNPR-sequence is t. Now, let $\mathcal{C}$ be a class of phylogenetic networks on X. We say that $\mathcal{C}$ is *connected* (under SNPR) if, for all $\mathcal{N}, \mathcal{N}' \in \mathcal{C}$, there is a SNPR-sequence connecting $\mathcal{N}$ and $\mathcal{N}'$ whereby every network in the sequence is in $\mathcal{C}$. If $\mathcal{C}$ is connected, then the SNPR-*distance* between two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ in $\mathcal{C}$, denoted $d_{\mathrm{SNPR}_\mathcal{C}}(\mathcal{N}, \mathcal{N}')$ or simply $d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{N}')$ if the context is clear, is the minimum length of a SNPR-sequence connecting $\mathcal{N}$ and $\mathcal{N}'$, where every network in the sequence is in $\mathcal{C}$. Furthermore, the *diameter* of the space of $\mathcal{C}$ (under SNPR) is the value

$$\max\{d_{\mathrm{SNPR}_\mathcal{C}}(\mathcal{N}, \mathcal{N}') : \mathcal{N}, \mathcal{N}' \in \mathcal{C}\}.$$

Proposition 3.2. *Under* SNPR*, the spaces of all phylogenetic networks, tree-child networks, reticulation-visible, and tree-based networks on X are connected. Moreover, the diameter of the space of all phylogenetic networks and tree-based networks on X is unbounded, while the diameters of each of the spaces of tree-child and reticulation-visible networks on X is $O(n)$, where $n = |X|$, and this is sharp.*

Proof. Let $\mathcal{N}$ and $\mathcal{N}'$ be two networks in $\mathcal{C}$, where $\mathcal{C}$ is one of the classes in the statement of the proposition. First suppose that $\mathcal{C}$ is the class of all phylogenetic networks on X. Let $(v_1, v_2, \ldots, v_k)$ be an ordering of the reticulations in $\mathcal{N}$ so that, for all distinct $i, j \in \{1, 2, \ldots, k\}$ with $i < j$, the reticulation v_i is not a descendant of v_j. Since $\mathcal{N}$ is acyclic, such an ordering exists. Now, taking this ordering, start with $\mathcal{N}$ and apply (II) to a reticulation edge incident with v_1, then apply (II) to a reticulation edge incident with v_2, and continue in this way for $v_3, v_4, \ldots, v_k$. After k applications, we obtain a phylogenetic network with no reticulations, that is, a rooted binary phylogenetic X-tree $\mathcal{T}$. Similarly, let $\mathcal{T}'$ be a rooted binary phylogenetic X-tree obtained from $\mathcal{N}'$ in an analogous way. Since the rSPR operation on trees induces a metric on the space of all rooted binary phylogenetic X-trees (Bordewich and Semple, 2004), there is an SNPR-sequence connecting $\mathcal{T}$ and

$\mathcal{T}'$. Thus, by Proposition 3.1, the proposition holds for when $\mathcal{C}$ is the class of all phylogenetic networks on X. If $\mathcal{C}$ is the classes of either tree-child, reticulation-visible, or tree-based networks on X, then the same approach works using a deletion ordering of $\mathcal{N}$ and $\mathcal{N}'$. Thus the spaces of tree-child, reticulation-visible, and tree-based networks on X are also connected.

For the proof of the second part of the proposition, the size of the leaf set of a phylogenetic network does not, in general, bound its total number of reticulations. However, it is shown in Cardona et al. (2009b) and Bordewich and Semple (2016) that tree-child and reticulation-visible networks on X have at most $n-1$ and $3n-2$ reticulations, respectively, and these bounds are sharp. The second part of the proposition now follows by noting that a single SNPR can remove at most one reticulation and each of the classes under consideration contains the class of rooted binary phylogenetic X-trees. □

The following corollary is an immediate consequence of Propositions 3.1 and 3.2.

Corollary 3.3. *The* SNPR*-distance is a metric on the classes of all phylogenetic networks, tree-child networks, reticulation-visible, and tree-based networks on X.*

4. Tree-Child Networks

In this section, we consider tree-child networks and start with a definition that generalizes the notion of a tree-child network. A phylogenetic network on X is *almost tree child* if either $\mathcal{N}$ is tree child or $\mathcal{N}$ has exactly one pair of parallel edges and the phylogenetic network obtained from $\mathcal{N}$ by deleting one of these edges and contracting the resulting degree-two vertices is tree child. Now, let $\mathcal{C}$ be the class of tree-child networks on X, and let $\mathcal{C}'$ be the class of almost tree-child networks on X. We say that $\mathcal{C}$ is *weakly connected* (under SNPR) if, for all $\mathcal{N}, \mathcal{N}' \in \mathcal{C}$, there is an SNPR-sequence connecting $\mathcal{N}$ and $\mathcal{N}'$ whereby every network in the sequence is in $\mathcal{C}'$. Note that if $\mathcal{N}$ is a tree-child network on X, then $\mathcal{N}$ has at most $|X|-1$ reticulations (Cardona et al., 2009b). The focus of this section is to prove the following theorem.

Theorem 4.1. *Let k be a fixed non-negative integer. If $k \leq |X|-2$, then, under* SNPR*, the space of tree-child networks on X with exactly k reticulations is connected. Otherwise, if $k = |X|-1$, then, under* SNPR*, the space of tree-child networks on X with exactly k reticulations is weakly connected.*

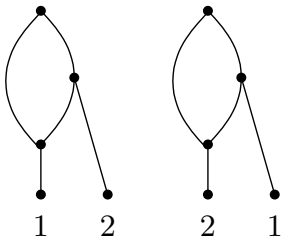

Figure 4: Two tree-child networks $\mathcal{N}$ and $\mathcal{N}'$ on $X = \{1, 2\}$ with exactly $|X| - 1$ reticulations. Note that there is no SNPR-sequence connecting $\mathcal{N}$ and $\mathcal{N}'$ such that every network in the sequence is tree child and has exactly $|X| - 1$ reticulations.

Moreover, in both of these spaces, the SNPR*-distance between any pair of tree-child networks on X is at most $O(n)$, where $n = |X|$.*

The discrepancy in Theorem 4.1 when $k = |X| - 1$ is that, under SNPR, the space of tree-child networks on X with exactly $|X| - 1$ reticulations may not be connected. For example, there is no SNPR-sequence in this space connecting the two tree-child networks shown in Fig. 4.

Loosely speaking, we prove Theorem 4.1 by showing that, for each tree-child network $\mathcal{N}$ on X, there is an appropriate SNPR-sequence connecting $\mathcal{N}$ and a particular type of tree-child network which we call a 'strict caterpillar network'.

Let $\mathcal{N}$ be a phylogenetic network. If $e = (u, v)$ is an edge in $\mathcal{N}$ that has no reticulations below v, then the subgraph containing v that is obtained from $\mathcal{N}$ by deleting e is called a *pendant subtree* of $\mathcal{N}$. We say that e *induces* a pendant subtree. Now suppose that $\mathcal{N}$ has no edges in parallel. We call $\mathcal{N}$ a *caterpillar network* if the following two properties are satisfied:

(i) the outgoing edge of each reticulation induces a pendant subtree of $\mathcal{N}$, and

(ii) there exists an ordering $(v_1, v_2, \ldots, v_k)$ of the reticulations such that there is a tree path starting at ρ whose first $2k + 1$ vertices are

$$\rho, p_1, q_1, p_2, q_2, \ldots, p_k, q_k,$$

where, for each i, the parents of v_i are p_i and q_i.

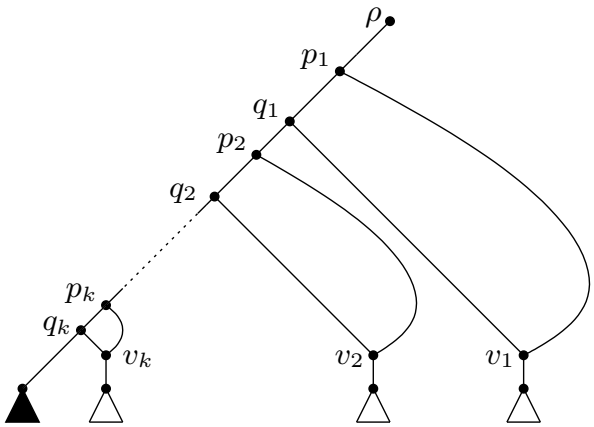

Figure 5: A caterpillar network $\mathcal{N}$ on k reticulations with reticulation ordering $(v_1, v_2, \ldots, v_k)$. Small triangles indicate pendant subtrees while the filled triangle corresponds to the tail of $\mathcal{N}$.

Observe that a caterpillar network is tree child. We call the ordering in (ii), the *reticulation ordering* in $\mathcal{N}$. Also, the pendant subtree induced by the outgoing edge of q_k that is not (q_k, v_k) is the *tail* while, for each $i \in \{1, 2, \ldots, k\}$, the pendant subtree induced by the outgoing edge of v_i is the *subtree below* v_i. A caterpillar network is *strict* if the subtree below each reticulation consists of a single leaf. To illustrate, a caterpillar network with reticulation ordering $(v_1, v_2, \ldots, v_k)$ is shown in Fig. 5.

Let $\mathcal{N}$ be a caterpillar network. Let v be a reticulation in $\mathcal{N}$ such that the subtree below v consists of at least two leaves, and let ℓ be one such leaf whose parent is u. Furthermore, let ℓ' be a leaf in the tail of $\mathcal{N}$. By the definition of a caterpillar network, note that ℓ' always exists. Now consider the caterpillar network $\mathcal{N}'$ obtained from $\mathcal{N}$ by applying (I) that deletes (u, ℓ), contracts u, subdivides the edge directed into ℓ' with a new vertex u', and adds the new edge (u', ℓ). We say that $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by *moving ℓ to the tail*. Note that $\mathcal{N}$ and $\mathcal{N}'$ have the same number of reticulations.

In the proof of the next lemma, we implicitly use the following characterisation of tree-child networks. The equivalence of (i) and (ii) is well known, while the equivalence of (i) and (iii) is shown in Semple (2016).

Proposition 4.2. *The following statements are equivalent for a phylogenetic network $\mathcal{N}$ on X with no parallel edges.*

(i) *$\mathcal{N}$ is a tree-child network.*

(ii) *Each non-leaf vertex in $\mathcal{N}$ has a child that is a tree vertex or a leaf.*

(iii) *No reticulation of $\mathcal{N}$ has a child reticulation, and no tree vertex of $\mathcal{N}$ has two child reticulations.*

Lemma 4.3. *Let k be a fixed non-negative integer, and let $\mathcal{N}$ be a tree-child network on X with exactly k reticulations. Then there is a* SNPR*-sequence*

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}'_1, \mathcal{N}_1, \mathcal{N}'_2, \mathcal{N}_2, \ldots, \mathcal{N}'_k, \mathcal{N}_k, \mathcal{N}_{k+1}, \mathcal{N}_{k+2}, \ldots, \mathcal{N}_t,$$

where $t \leq n + k$, with the following properties:

(i) *each network in the sequence is tree child with exactly k reticulations;*

(ii) *$\mathcal{N}_i$ is a caterpillar network for all $i \in \{k, k+1, \ldots, t\}$; and*

(iii) *$\mathcal{N}_t$ is a strict caterpillar network.*

Proof. Let w be the unique child of ρ in $\mathcal{N}$. Furthermore, let $(v_1, v_2, \ldots, v_k)$ be an arbitrary ordering of the reticulations in $\mathcal{N}$ and note that, for all i, the parents of v_i, say u_i and u'_i, are both tree vertices.

We begin by describing algorithmically a SNPR-sequence connecting $\mathcal{N}$ with a caterpillar network on X. Setting $\mathcal{N} = \mathcal{N}_0$, $i = 1$, and $q_0 = \rho$, repeat the following 2-step process k times:

1. Apply (I) to (u_i, v_i) that results in subdividing the non-reticulation edge directed out of q_{i-1} with a new vertex p_i, and adding a new edge (p_i, v_i). Let $\mathcal{N}'_i$ denote the resulting phylogenetic network on X.

2. Apply (I) to (u'_i, v_i) with $u'_i \neq p_i$ that results in subdividing the edge directed out of p_i that is not (p_i, v_i) with a new vertex q_i, and adding a new edge (q_i, v_i). Let $\mathcal{N}_i$ denote the resulting phylogenetic network on X. If $i = k$ stop; otherwise, increment i by one and repeat.

Furthermore, if, for some $i \in \{1, 2, \ldots, k\}$, we have $w = u_i$ or $w = u'_i$, then, after contracting w, rename the child tree vertex of w as w in the resulting network.

We first show by induction that, for all $i \in \{1, 2, \ldots, k\}$, both $\mathcal{N}'_i$ and $\mathcal{N}_i$ are tree-child networks with exactly k reticulations. By definition, $\mathcal{N}_0$

is a tree-child network with exactly k reticulations. Now suppose that $\mathcal{N}_{i-1}$ is a tree-child network with exactly k reticulations, where $i \geq 1$. By construction, $\mathcal{N}_i'$ has exactly k reticulations. Furthermore, as $\mathcal{N}_{i-1}$ is tree child, u_i has exactly one child reticulation, namely v_i. Thus, deleting (u_i, v_i) and contracting u_i does not result in any pair of parallel edges, and so $\mathcal{N}_i'$ has no edges in parallel. Lastly, as $\mathcal{N}_{i-1}$ is tree child, the only plausible non-leaf vertices in $\mathcal{N}_i'$ that may not have a child that is a tree vertex or a leaf are p_i and the parent, say t_i of u_i in $\mathcal{N}_{i-1}$. But w, a tree vertex, is a child of p_i in $\mathcal{N}_i'$. Also, the child of u_i in $\mathcal{N}_{i-1}$ that is not v_i is either a tree vertex or a leaf, and this child is now the child of t_i in $\mathcal{N}_i'$. Hence $\mathcal{N}_i'$ is a tree-child network with exactly k reticulations. Using $\mathcal{N}_i'$ and a similar argument shows that $\mathcal{N}_i$ is also a tree-child network with exactly k reticulations.

We now show that $\mathcal{N}_k$ is a caterpillar network. By construction, for all $i \in \{1, 2, \ldots, k\}$, there is no reticulation below v_i in $\mathcal{N}_k$ and so the outgoing edge of v_i induces a pendant subtree in $\mathcal{N}_k$. Further, since $\mathcal{N}$ is tree child, there exists a tree path $\rho, w, \ldots, \ell$ in $\mathcal{N}$ for some leaf ℓ and so, again by construction,

$$\rho, p_1, q_1, p_2, q_2, \ldots, q_k, q_k, \ldots, \ell$$

is a tree path in $\mathcal{N}_k$. It follows that $\mathcal{N}_k$ is a caterpillar network.

To complete the proof, we describe an appropriate SNPR-sequence connecting $\mathcal{N}_k$ and a strict caterpillar network on X. If $\mathcal{N}_k$ is strict, we are done; otherwise, there exists a reticulation v_i in $\mathcal{N}_k$ such that the subtree below v_i has at least two leaves. Let ℓ' be one such leaf. By moving ℓ' to the tail of $\mathcal{N}_k$, we obtain a caterpillar network $\mathcal{N}_{k+1}$ on X with exactly k reticulations whose subtree below v_i has one leaf less than the subtree below v_i in $\mathcal{N}_k$. We can repeatedly apply this process at most $n - k$ times to eventually obtain a strict caterpillar network $\mathcal{N}_t$ on X with exactly k reticulations by moving all but one leaf of each subtree below a reticulation to the tail, thereby completing the proof of the lemma. □

Lemma 4.4. *Let k be a fixed non-negative integer, let $\mathcal{N}$ and $\mathcal{N}'$ be strict caterpillar networks on X with exactly k reticulations, and let $n = |X|$. Then there is a* SNPR*-sequence connecting $\mathcal{N}$ and $\mathcal{N}'$ of length at most $2(n+3k-1)$ such that each network in the sequence is almost tree child and has exactly k reticulations. Moreover, if $k \leq |X| - 2$, then there is a* SNPR*-sequence connecting $\mathcal{N}$ and $\mathcal{N}'$ of length at most $2(n + 3k - 1)$ such that each network in the sequence is tree child and has exactly k reticulations.*

Proof. First observe that if $k \leq |X| - 2$, then the tail of a strict caterpillar network on X contains at least two leaves; otherwise, it contains exactly one leaf. This observation is used throughout the proof. Let $(\ell_1, \ell_2, \ldots, \ell_n)$ be an ordering of the elements in X. Now, up to isomorphism, let $\mathcal{M}$ be the strict caterpillar network on X with reticulation ordering $(v_1, v_2, \ldots, v_k)$ such that, for all i, the leaf below v_i is ℓ_i and whose tail is the caterpillar tree

$$(\ell_n, \ell_{n-1}, \ldots, \ell_{k+1});$$

that is the rooted binary phylogenetic tree in which ℓ_n and ℓ_{n-1} have the same parent and, for each $j \in \{n-2, n-3, \ldots, k+1\}$, the parent of ℓ_{j+1} is a child of the parent of ℓ_j. To prove the lemma it suffices to show, by Proposition 3.1, that there is a SNPR-sequence connecting $\mathcal{N}$ and $\mathcal{M}$ of length at most $n + 3k - 1$ such that each network in the sequence is almost tree child with exactly k reticulations if $k = |X| - 1$ and, otherwise, each network in the sequence is tree child with exactly k reticulations.

We begin by describing a SNPR-sequence $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$, where each network has leaf set X and exactly k reticulations, $\mathcal{N}_1$ and $\mathcal{N}_3$ are both strict caterpillar networks, and $\mathcal{N}_2$ is almost tree child if $k = |X| - 1$ and tree child if $k \leq |X| - 2$. The sequence allows us to interchange a leaf in the tail of $\mathcal{N}_1$ with a leaf below any one of its reticulations.

Let $\mathcal{N}_1$ be a strict caterpillar network on X with exactly k reticulations. Let ℓ be a leaf in the tail of $\mathcal{N}_1$ and let ℓ' be the leaf below a reticulation v' in $\mathcal{N}_1$. Suppose that (u, ℓ) is the edge in $\mathcal{N}_1$ that is directed into ℓ. Let $\mathcal{N}_2$ be the phylogenetic network on X obtained by applying (I) to (u, ℓ) that results in subdividing (v', ℓ') with a new vertex u', and adding a new edge (u', ℓ). Since $\mathcal{N}_1$ is a strict caterpillar network on X with exactly k reticulations, $\mathcal{N}_2$ is a caterpillar network on X with exactly k reticulations unless $k = |X| - 1$. In the exceptional case, the last reticulation in the reticulation ordering of $\mathcal{N}_1$ is the end vertex of two edges in parallel in $\mathcal{N}_2$. However, it is easily checked that there are no other parallel edges in $\mathcal{N}_2$, and that deleting one of the these edges and contracting the resulting degree-two vertices gives a caterpillar network on X with $k - 1$ reticulations. Thus $\mathcal{N}_2$ is almost tree child.

To obtain $\mathcal{N}_3$, move ℓ' to the tail if $\mathcal{N}_2$ is a caterpillar network; otherwise, apply (I) to (u', ℓ') that results in subdividing one of the parallel edges with a new vertex u'', and adding a new edge (u'', ℓ'). Regardless of the construction, it is easily checked that $\mathcal{N}_3$ is a strict caterpillar network on X with exactly k reticulations.

Now we can repeatedly apply the above SNPR-sequence of length two to obtain a SNPR-sequence connecting $\mathcal{N}$ and, up to isomorphism, a strict caterpillar network $\mathcal{M}'$ on X with reticulation ordering $(v'_1, v'_2, \ldots, v'_k)$ such that, for all i, the leaf below v'_i is ℓ_i. This can be done with a sequence of length at most $4k$. Moreover, each network in the sequence is almost tree child with exactly k reticulations if $k = |X| - 1$ and, otherwise, each network in the sequence is tree child with exactly k reticulations. Extending this sequence, we can apply (I) at most $n - (k + 1)$ times, with each application subdividing an edge within the tail and so each network in the sequence is a strict caterpillar network on X with exactly k reticulations, to eventually obtain $\mathcal{M}$. This completes the proof of the lemma. □

We now combine Lemmas 4.3 and 4.4 to prove Theorem 4.1.

Proof of Theorem 4.1. Let $\mathcal{N}$ and $\mathcal{N}'$ be two tree-child networks on X with exactly k reticulations. Let $\mathcal{M}$ be a strict caterpillar network on X with exactly k reticulations. By Lemmas 4.3 and 4.4, and the fact that $k \leq n$, there is a SNPR-sequence connecting $\mathcal{N}$ and $\mathcal{M}$ of length at most $O(n)$ such that each network in the sequence is almost tree child with exactly k reticulations if $k = |X| - 1$ and, otherwise, tree child with exactly k reticulations. An analogous sequence connects $\mathcal{N}'$ and $\mathcal{M}$. The theorem now follows from Proposition 3.1. □

5. Tree-Based and Reticulation-Visible Networks

Tree-based and reticulation-visible networks were defined as subclasses of phylogenetic networks with only the former class allowing for parallel edges. However, there is no reason not to include parallel edges in the definition of reticulation-visible networks as well. For the purposes of this section, we take this viewpoint. In particular, we will allow tree-based and reticulation-visible networks to have edges in parallel. The main results of this section are the next theorem for tree-based networks and the analogous result for reticulation-visible networks at the end of this section.

Theorem 5.1. *Let k be a fixed non-negative integer. Then, under* SNPR*, the space of tree-based networks on X with exactly k reticulations is connected. Moreover, if $\mathcal{T}$ is a fixed rooted binary phylogenetic X-tree, then, under* SNPR*, the space of phylogenetic networks on X with base tree $\mathcal{T}$ and exactly k reticulations is connected. Moreover, the diameter of both of these spaces is at most $O(kn)$, where $n = |X|$.*

Note that, unlike tree-child networks, the size of the leaf set does not bound the total number of vertices, and therefore the number of reticulations, of a tree-based network $\mathcal{N}$ even if $\mathcal{N}$ has no parallel edges. Thus, as a consequence, quite a different approach is used to prove Theorem 5.1 in comparison with that used to prove Theorem 4.1.

We begin with two lemmas. A reticulation v in a phylogenetic network is said to be *in parallel* if the two reticulation edges incident with v are a pair of parallel edges.

Lemma 5.2. *Let k be a fixed non-negative integer, and let $\mathcal{N}$ be a tree-based network on X with exactly k reticulations. Suppose that $\mathcal{T}$ is a base tree for $\mathcal{N}$. Then there is a* SNPR*-sequence*

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_s,$$

where $s \leq k$, with the following properties:

(i) *for each $i \in \{0, 1, 2, \ldots, s\}$, we have $\mathcal{N}_i$ is a tree-based network with base tree $\mathcal{T}$ and exactly k reticulations; and*

(ii) *each reticulation in $\mathcal{N}_s$ is in parallel.*

Proof. Let $\mathcal{S}$ be an embedding of $\mathcal{T}$ in $\mathcal{N}$ with vertex set $V(\mathcal{S})$ and edge set $E(\mathcal{S})$. Let v be a reticulation in $\mathcal{N}$ that is not in parallel, and let e and f be the reticulation edges incident with v. Furthermore, let u_e and u_f be the end vertices of e and f not equal to v, respectively. As v is not in parallel, $u_e \neq u_f$. Since $\mathcal{S}$ is a base tree for $\mathcal{N}$, precisely one of e and f is an edge in $\mathcal{S}$. Without loss of generality, we may assume that e is an edge in $\mathcal{S}$. Then u_f is a tree vertex; otherwise, u_f is not a vertex in $\mathcal{S}$. Let g_1 be the edge directed into u_f and let g_2 be the edge directed out of u_f that is not incident with v.

Now consider the phylogenetic network $\mathcal{N}_1$ on X obtained from $\mathcal{N}$ by applying (I) to f that results in contracting u_f to create a new edge g, subdividing e with a new vertex u'_f, and adding a new edge in parallel with (u'_f, v). Now, in $\mathcal{N}_1$, the reticulation v is in parallel. Moreover, as $\mathcal{N}$ is a tree-based network with exactly k reticulations and $\mathcal{S}$ is a base tree for $\mathcal{N}$, it follows that $\mathcal{N}_1$ is a tree-based network with exactly k reticulations and $\mathcal{S}_1$ with vertex set $(V(\mathcal{S}) - \{u_f\}) \cup \{u'_f\}$ and edge set $(E(\mathcal{S}) - \{e, g_1, g_2\}) \cup$

$\{(u_e, u'_f), (u'_f, v), g\}$, is a base tree for $\mathcal{N}_1$. Moreover, as $\mathcal{S}$ is an embedding of $\mathcal{T}$ in $\mathcal{N}$, it is easily seen that $\mathcal{S}_1$ is an embedding of $\mathcal{T}$ in $\mathcal{N}_1$.

Choosing a reticulation that is not in parallel in $\mathcal{N}_1$, and repeatedly applying this iterative process beginning with $\mathcal{N}_1$ and $\mathcal{S}_1$, we eventually obtain, for some $s \leq k$, a SNPR-sequence

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_s,$$

where $\mathcal{N}_i$ is a tree-based network with a base tree $\mathcal{T}$ and exactly k reticulations for all i, and $\mathcal{N}_s$ has the additional property that each reticulation is in parallel. □

Let $\mathcal{N}$ be a phylogenetic network on X. For $k \geq 0$, consider the phylogenetic network $\mathcal{N}'$ on X obtained from $\mathcal{N}$ by subdividing the unique edge that is directed out of ρ with $2k$ vertices

$$u_1, v_1, u_2, v_2, u_3, \ldots, u_k, v_k$$

such that $\rho, u_1, v_1, u_2, v_2, u_3, \ldots, u_k, v_k$ is a directed path in $\mathcal{N}'$ and then, for each $i \in \{1, 2, \ldots, k\}$, adding an edge in parallel with (u_i, v_i). We say that the reticulations $v_1, v_2, \ldots, v_k$ are in *series at* ρ in $\mathcal{N}'$.

Lemma 5.3. *Let k be a fixed non-negative integer, and let $\mathcal{N}$ be a tree-based network on X with exactly k reticulations each of which is in parallel. Let $n = |X|$. Suppose that $\mathcal{T}$ is a base tree for $\mathcal{N}$. Then there is a* SNPR*-sequence*

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_t,$$

where t is at most $O(kn)$, with the following properties:

(i) *for each $i \in \{0, 1, 2, \ldots, t\}$, we have $\mathcal{N}_i$ is a tree-based network with base tree $\mathcal{T}$ and exactly k reticulations; and*

(ii) *each reticulation in $\mathcal{N}_t$ is in series at ρ.*

Proof. Let $\mathcal{S}$ be an embedding of $\mathcal{T}$ in $\mathcal{N}$, and let $V(\mathcal{S})$ and $E(\mathcal{S})$ be the vertex set and edge set of $\mathcal{S}$, respectively. If each reticulation is already in series at ρ, we are done. Therefore assume that not all reticulations in $\mathcal{N}$ are in series at ρ. Recalling that each reticulation in $\mathcal{N}$ is in parallel, it is easily checked that there exists a reticulation, v say, in $\mathcal{N}$ that is not in series at

ρ such that each directed path from ρ to v only traverses reticulations that are in series at ρ, tree vertices, as well as ρ and v themselves. Note that the unique grandparent, w say, of v is a degree-3 vertex that is incident with three tree edges. Now, let e and f be the reticulation edges directed into v. Without loss of generality, we may assume that f is not contained in $E(\mathcal{S})$. Furthermore, let u be the unique parent of v. By the choice of v, the parent of u is w. Let $g = (w, y)$ be the edge directed out of w with $y \neq u$. We next apply (I) twice that, taken together, intuitively move v above w.

For the first application of (I), let $\mathcal{N}_1$ be the phylogenetic network on X obtained from $\mathcal{N}$ by deleting f, contracting u and thereby creating a new edge (w, v), subdividing g with a new vertex u', and adding a new edge (u', v). Clearly, as $\mathcal{N}$ has exactly k reticulations so does $\mathcal{N}_1$. Furthermore, as $\mathcal{N}$ is a tree-based network with base tree $\mathcal{S}$, it follows that $\mathcal{N}_1$ is a tree-based network and $\mathcal{S}_1$ with vertex set $(V(\mathcal{S}) - \{u\}) \cup \{u'\}$ and edge set $(E(\mathcal{S}) - \{(w, u), e, g\}) \cup \{(w, v), (w, u'), (u', y)\}$ is a base tree for $\mathcal{N}_1$. Also, since $\mathcal{S}$ is an embedding of $\mathcal{T}$ in $\mathcal{N}$, it is easily checked that $\mathcal{S}_1$ is an embedding of $\mathcal{T}$ in $\mathcal{N}_1$.

For the second application of (I), let $\mathcal{N}_2$ be the phylogenetic network on X obtained from $\mathcal{N}_1$ by deleting (u', y), contracting u' and thereby creating a new edge (w, v), subdividing the edge that is directed out of v with a new vertex x, and adding a new edge (x, y). As (w, v), (w, u'), and (u', v) are edges in $\mathcal{N}_1$, we have v being in parallel in $\mathcal{N}_2$. Again, as at the end of the last paragraph, $\mathcal{N}_2$ is a tree-based network with base tree $\mathcal{T}$ and exactly k reticulations.

Now, in comparison to $\mathcal{N}$, the reticulations in series in $\mathcal{N}$ are still in series in $\mathcal{N}_2$, but the number of edges on the shortest path from ρ to v has decreased by one. Since there are no reticulations on any directed path from ρ to v except for (possibly) reticulations in series at ρ, it follows that at most $O(n)$ operations are sufficient to put v in series at ρ. Hence, by repeatedly applying this 2-step process, we can eventually obtain, for some t of at most $O(kn)$, a SNPR-sequence

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_2, \ldots, \mathcal{N}_t,$$

where $\mathcal{N}_i$ is a tree-based network with base tree $\mathcal{T}$ and exactly k reticulations for all i, and $\mathcal{N}_t$ has the additional property that each reticulation is in series at ρ. This completes the proof of the lemma. □

Proof of Theorem 5.1. Let $\mathcal{N}$ be a tree-based network on X with exactly k

reticulations. Let $\mathcal{T}$ be a base tree for $\mathcal{N}$. By Lemmas 5.2 and 5.3, there is a SNPR-sequence of length at most $O(kn)$ connecting $\mathcal{N}$ and a tree-based network $\mathcal{M}$ with base tree $\mathcal{T}$ and exactly k reticulations each of which is in series at ρ such that each network in the sequence is tree based with base tree $\mathcal{T}$ and exactly k reticulations. Observe that, up to isomorphism, $\mathcal{M}$ is unique.

Now let $\mathcal{N}'$ be a tree-based network on X with exactly k reticulations. If $\mathcal{T}$ is a base tree for $\mathcal{N}'$, then, as in the last paragraph, there is a SNPR-sequence of length at most $O(kn)$ connecting $\mathcal{N}'$ and $\mathcal{M}$. It now follows by Proposition 3.1 that, for a non-negative integer k and a rooted binary phylogenetic X-tree $\mathcal{T}$, the space of tree-child networks on X with base tree $\mathcal{T}$ and exactly k reticulations is connected under SNPR and the diameter of this space is at most $O(kn)$.

To complete the proof of the first part of the theorem, let $\mathcal{T}'$ be a base tree for $\mathcal{N}'$. By Lemmas 5.2 and 5.3, there is a SNPR-sequence of length at most $O(kn)$ connecting $\mathcal{N}'$ and a tree-based network $\mathcal{M}'$ with base tree $\mathcal{T}'$ and exactly k reticulations each of which is in series at ρ such that each network in the sequence has exactly k reticulations. Consider $\mathcal{M}$ and $\mathcal{M}'$. Ignoring the reticulations in series at ρ, we can apply a sequence of at most n operations of type (I) (essentially rSPR operations) to connect $\mathcal{M}$ and $\mathcal{M}'$. The first part of the theorem now follows from Proposition 3.1. □

We end this section with a connectedness result for reticulation-visible networks. It is easily checked that Lemmas 5.2 and 5.3 hold with "tree-based networks" replaced by "reticulation-visible networks". The same proofs work as reticulation-visible networks are tree based. However, one does have to additionally check that, for each network in the SNPR-sequence, all reticulations are visible. The proof of the next theorem is similar to that used to prove Theorem 5.1 and omitted. Note that if $\mathcal{N}$ is a reticulation-visible network with n leaves and without parallel edges, then $\mathcal{N}$ has at most $3n-3$ reticulations (Bordewich and Semple, 2016).

Theorem 5.4. *Let k be a fixed non-negative integer. Then, under* SNPR*, the space of reticulation-visible networks on X with exactly k reticulations is connected. Moreover, the* SNPR*-distance between a pair of reticulation-visible networks on X without parallel edges is at most $O(n^2)$, while for a pair of arbitrary reticulation-visible networks this distance is at most $O(kn)$, where $n = |X|$.*

6. Networks that Display a Given Set of Trees

Let $\mathcal{P}$ be a given set of rooted binary phylogenetic X-trees. In this section, we show that the class of phylogenetic networks on X that display each tree in $\mathcal{P}$, that is, the class of phylogenetic networks on X that *display* $\mathcal{P}$, is connected.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree, and let $\mathcal{N}$ be a phylogenetic network on X. Obtain a phylogenetic network $\mathcal{N}'$ from $\mathcal{N}$ by subdividing (ρ, w) with a new vertex u, joining the root of $\mathcal{T}$ with u via a new edge and, for each leaf ℓ in $\mathcal{T}$, subdividing the pendant edge of $\mathcal{N}$ ending in ℓ with a new vertex v_ℓ and identifying the vertex labeled ℓ in $\mathcal{T}$ with v_ℓ. We say that $\mathcal{N}'$ *extends* $\mathcal{N}$ by $\mathcal{T}$.

Lemma 6.1. *Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree, and let $\mathcal{N}$ be a phylogenetic network on X. Let $n = |X|$. Then there is a* SNPR*-sequence*

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_n$$

such that the following hold:

(i) *If $\mathcal{T}'$ is a rooted binary phylogenetic X-tree displayed by $\mathcal{N}$, then, for all $i \in \{0, 1, 2, \ldots, n\}$, we have that $\mathcal{N}_i$ displays $\mathcal{T}'$.*

(ii) *The phylogenetic network $\mathcal{N}_n$ extends $\mathcal{N}$ by $\mathcal{T}$.*

Proof. Let $X = \{\ell_1, \ell_2, \ldots, \ell_n\}$. Consider the following iterative process:

1. Apply (III) to the edge, e_1 say, incident with ℓ_1 that results in subdividing e_1 with a new vertex v_1, subdividing the edge incident with ρ with a new vertex u_1, and adding a new edge (u_1, v_1). Let $\mathcal{N}_1$ denote the resulting phylogenetic network on X. Set $E_1 = \{(u_1, v_1)\}$ and $i = 2$.

2. Apply (III) to the edge, e_i say, incident with ℓ_i that results in subdividing e_i with a new vertex v_i, subdividing one of the edges in E_{i-1}, say (w, w'), with a new vertex u_i such that once the new edge (u_i, v_i) is added the edges in

   $$(E_{i-1} - \{(w, w')\}) \cup \{(w, u_i), (u_i, w'), (u_i, v_i)\} \cup \{(v_1, \ell_1), (v_2, \ell_2), \ldots, (v_i, \ell_i)\}$$

   and the leaves $\ell_1, \ell_2, \ldots, \ell_i$ form an embedding of $\mathcal{T}|\{\ell_1, \ell_2, \ldots, \ell_i\}$ in the resulting network $\mathcal{N}_i$. If $i = n$, stop; otherwise, increment i by one, set $E_i = (E_{i-1} - \{(w, w')\}) \cup \{(w, u_i), (u_i, w'), (u_i, v_i)\}$, and repeat this step.

By construction, $\mathcal{N}_n$ extends $\mathcal{N}$ by $\mathcal{T}$. Furthermore, as no edge is ever deleted in this process, for all $i \in \{0, 1, 2, \ldots, n\}$, we have $\mathcal{N}_i$ displaying every rooted binary phylogenetic X-tree displayed by $\mathcal{N}$. This completes the proof of the lemma. □

Now let $\mathcal{P}$ be a set of rooted binary phylogenetic X-trees, and let $\mathcal{N}$ be a phylogenetic network on X that displays $\mathcal{P}$. Suppose that $O = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_t)$ is an ordering on the trees in $\mathcal{P}$. Setting $\mathcal{N} = \mathcal{N}_0$, for each $i \in \{1, 2, \ldots, t\}$, let $\mathcal{N}_i$ be the phylogenetic network on X that extends $\mathcal{N}_{i-1}$ by $\mathcal{T}_i$. We say that $\mathcal{N}_t$ *extends* $\mathcal{N}$ by O. For a fixed O, observe that, up to isomorphism, $\mathcal{N}_t$ is unique.

Theorem 6.2. *Let $\mathcal{P}$ be a set of rooted binary phylogenetic X-trees. Then, under* SNPR*, the space of all phylogenetic networks on X that display $\mathcal{P}$ is connected. Moreover, the distance between any two networks $\mathcal{N}$ and $\mathcal{N}'$ in this space is at most*

$$2(t+1)n + k + k',$$

where $t = |\mathcal{P}|$, $n = |X|$, and k and k' are the number of reticulations in $\mathcal{N}$ and $\mathcal{N}'$, respectively.

Proof. Let $n = |X|$, and let $O = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_t)$ be an ordering on the trees in $\mathcal{P}$. Furthermore, let $\mathcal{N}$ be a phylogenetic network on X that displays $\mathcal{P}$, and, up to isomorphism, let $\mathcal{N}_t$ be the extension of $\mathcal{N}$ by O. By repeated applications of Lemma 6.1, there exists a SNPR-sequence of length tn connecting $\mathcal{N}$ and $\mathcal{N}_t$ such that each network in the sequence displays $\mathcal{P}$. Also, by definition, if $e = (u, v)$ is an edge in $\mathcal{N}$ that is not directed into a leaf, then $e = (u, v)$ is an edge in $\mathcal{N}_t$.

Now, let $(v_1, v_2, \ldots, v_k)$ be an ordering on the reticulations in $\mathcal{N}$ such that, if v_j is a descendant of v_i in $\mathcal{N}$, then v_i precedes v_j. As $\mathcal{N}$ is acyclic, such an ordering exists. Under this ordering, take $\mathcal{N}_t$ and apply (II) to a reticulation edge directed into v_1, then apply (II) to a reticulation edge directed into v_2, and continue in this way for $v_3, v_4, \ldots, v_k$. Note that, because of O, these operations are allowable. Also, observe that after each application, the resulting network displays $\mathcal{P}$. Let $\mathcal{N}_{t+k}$ denote the phylogenetic network on X obtained at the end of these k applications, and note that we have now constructed a SNPR-sequence of length $tn+k$ connecting $\mathcal{N}$ and $\mathcal{N}_{t+k}$, where each network in the sequence displays $\mathcal{P}$.

Consider $\mathcal{N}_{t+k}$, and say $X = \{\ell_1, \ell_2, \ldots, \ell_n\}$. For all $i \in \{1, 2, \ldots, n\}$, there is a directed path

$$w_i^1, w_i^2, \ldots, w_i^t, \ell_i$$

in $\mathcal{N}_{t+k}$ such that each of $w_i^1, w_i^2, \ldots, w_i^t$ is a reticulation that was added when extending $\mathcal{N}$ by O to obtain $\mathcal{N}_t$. For each i, exactly one of the two parents of w_i^1, say p_i, is a vertex in $\mathcal{N}$. Now starting with $\mathcal{N}_{t+k}$, apply (II) to (p_1, w_1^1), then apply (II) to (p_2, w_2^1), and repeat for each of $(p_3, w_3^1), (p_4, w_4^1), \ldots, (p_n, w_n^1)$. After each application, the resulting network displays $\mathcal{P}$. Let $\mathcal{M}$ denote the phylogenetic network on X obtained at the end of these n applications of (II). It is easily seen that, up to isomorphism, $\mathcal{M}$ is the phylogenetic network on X that is the extension of $\mathcal{T}_1$ by $(\mathcal{T}_2, \mathcal{T}_3, \ldots, \mathcal{T}_t)$. Moreover, we have constructed a SNPR-sequence of length $(tn + k) + n = (t + 1)n + k$ connecting $\mathcal{N}$ and $\mathcal{M}$, where each network in the sequence displays $\mathcal{P}$.

We complete the proof by noting that if $\mathcal{N}'$ is a phylogenetic network on X that displays $\mathcal{P}$, then there is a SNPR-sequence of length $(t + 1)n + k'$ connecting $\mathcal{N}'$ and $\mathcal{M}$, where k' is the number of reticulations in $\mathcal{N}'$. The proof now follows from Proposition 3.1. □

7. Computing the SNPR-Distance

In this section, we show that computing the SNPR-distance between an arbitrary pair of phylogenetic networks on X is NP-hard, as well as establishing a related upper bound and analysing the relationship between the SNPR- and rSPR-distances. We begin with a proposition whose proof makes use of the notion of agreement forests, a well known and valuable tool in the context of the rSPR operation for rooted binary phylogenetic trees.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. As with phylogenetic networks, we view the roots ρ of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex of in-degree zero and out-degree one adjacent to the original root of $\mathcal{T}$ and $\mathcal{T}'$, respectively. A *forest* is a partition $\{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ of $X \cup \{\rho\}$ with $\rho \in \mathcal{L}_\rho$ such that the trees in

$$\big\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\big\}$$

are vertex-disjoint subtrees of $\mathcal{T}$. Furthermore, a forest $\{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ for $\mathcal{T}$ and $\mathcal{T}'$ is an *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ if, for all $i \in \{\rho, 1, 2, \ldots, k\}$, we have

$$\mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i.$$

A *maximum agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is an agreement forest in which k is minimised. The minimum possible value for k is denoted $m(\mathcal{T}, \mathcal{T}')$. It is shown in Bordewich and Semple (2004) that

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}'). \tag{1}$$

Furthermore, it is also shown in Bordewich and Semple (2004) that computing the rSPR-distance between an arbitrary pair of rooted binary phylogenetic X-trees is NP-hard. We utilise these two results in the proofs of the next proposition and the subsequent theorem.

Proposition 7.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{SNPR}}(\mathcal{T}, \mathcal{T}').$$

Furthermore, if there is a SNPR*-sequence connecting $\mathcal{T}$ and $\mathcal{T}'$ of length k and $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = k$, then every network in the sequence is a rooted binary phylogenetic X-tree.*

Proof. Since a rSPR operation is a certain instance of a SNPR operation, we have

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \geq d_{\mathrm{SNPR}}(\mathcal{T}, \mathcal{T}').$$

To establish the converse, let

$$\mathcal{T} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k = \mathcal{T}'$$

be a SNPR-sequence connecting $\mathcal{T}$ and $\mathcal{T}'$, and consider the following 2-colouring of the edges of each network in the sequence. Colour each of the edges in $\mathcal{T} = \mathcal{N}_0$ blue. For each $i \in \{1, 2, \ldots, k\}$, we preserve the edge colouring of $\mathcal{N}_{i-1}$ to colour the edges in $\mathcal{N}_i$ except for those edges changed by the SNPR operation applied to $\mathcal{N}_{i-1}$ to obtain $\mathcal{N}_i$. In particular, if a vertex is contracted, the resulting edge is coloured blue if both edges incident with the contracted vertex immediately prior to the contraction are blue, otherwise the resulting edge is coloured red. Furthermore, the resulting two edges of a subdivision are coloured the same colour as that of the edge being subdivided, and if a new edge is added, this new edge is coloured red.

Ignoring empty sets, for all $i \in \{0, 1, 2, \ldots, k\}$, let $\mathcal{F}_i$ be the partition of $X \cup \{\rho\}$ induced by the components of the directed graph obtained from $\mathcal{N}_i$ by deleting each red edge. Clearly, $\mathcal{F}_k$ is a forest for $\mathcal{T}'$. We now show that,

for all i, the partition $\mathcal{F}_i$ is a forest for $\mathcal{T}$ and consists of at most $i+1$ parts. The proof is by induction on i. Evidently, $\mathcal{F}_0$ consists of one part and is a forest for $\mathcal{T}$. Now suppose that $i \geq 1$ and $\mathcal{F}_{i-1}$ is a forest for $\mathcal{T}$ and consists of at most i parts. First assume that $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by applying either (I) or (II), and let e denote the edge deleted in this operation. Then $\mathcal{F}_i$ is the partition of $X \cup \{\rho\}$ induced by the components of the directed graph obtained from $\mathcal{N}_{i-1}$ by deleting each red edge as well as e. Since $\mathcal{F}_{i-1}$ is a forest for $\mathcal{T}$ and consists of at most i parts, it follows by induction that $\mathcal{F}_i$ is a forest for $\mathcal{T}$ and consist of at most $i+1$ parts. Now assume that $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by applying (III). Then, it is easily checked that $\mathcal{F}_i = \mathcal{F}_{i-1}$ and so, by induction, $\mathcal{F}_i$ is a forest for $\mathcal{T}$ and consists of at most $i+1$ parts.

Since $\mathcal{F}_k$ is a forest for $\mathcal{T}$ and $\mathcal{T}'$, it follows that $\mathcal{F}_k$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Also, by the induction argument, $|\mathcal{F}_k| \leq k+1$ and so, by (1),

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq k.$$

To complete the proof of the proposition, suppose that there is a SNPR-sequence connecting $\mathcal{T}$ and $\mathcal{T}'$ of length k and $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = k$. We next show that, for all $i \in \{1, 2, \ldots, k\}$, the phylogenetic network $\mathcal{N}_i$ is a rooted binary phylogenetic X-tree, i.e. $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by applying (I) without creating a pair of parallel edges. Assume the contrary. Then, for some $j \in \{1, 2, \ldots, k\}$, the phylogenetic network $\mathcal{N}_j$ is obtained from $\mathcal{N}_{j-1}$ by applying (II). In turn, this implies that, for some $i \in \{1, 2, \ldots, j-1\}$, the phylogenetic network $\mathcal{N}_i$ has been obtained from $\mathcal{N}_{i-1}$ by applying (III). But then, $\mathcal{F}_i = \mathcal{F}_{i-1}$ and so

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq k-1;$$

a contradiction. This completes the proof of the proposition. □

The next theorem follows immediately from this result and the earlier-mentioned fact that computing the rSPR-distance between $\mathcal{T}$ and $\mathcal{T}'$ is NP-hard.

Theorem 7.2. *Computing the* SNPR*-distance between an arbitrary pair of phylogenetic networks on X is* NP*-hard.*

We next establish an upper bound on the SNPR-distance between two phylogenetic networks on X. We begin with two lemmas. The first is an

immediate consequence of the fact that a rooted binary phylogenetic tree has no vertex with in-degree two and a single SNPR can delete at most one reticulation edge.

Lemma 7.3. *Let $\mathcal{N}$ be a phylogenetic network on X, and let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. Then*

$$d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{T}) \geq k,$$

where k is the number of reticulations in $\mathcal{N}$.

Lemma 7.4. *Let $\mathcal{N}$ be a phylogenetic network on X, and let $\mathcal{T}$ be a rooted binary phylogenetic X-tree. If $\mathcal{N}$ displays $\mathcal{T}$, then*

$$d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{T}) = k,$$

where k is the number of reticulations in $\mathcal{N}$.

Proof. Suppose that $\mathcal{N}$ displays $\mathcal{T}$. By Lemma 7.3, it suffices to show that $d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{T}) \leq k$. The proof is by induction on k. If $k = 0$, then $\mathcal{N} \cong \mathcal{T}$, and so $d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{T}) = 0$, that is,

$$d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{T}) \leq 0.$$

Now suppose that $k \geq 1$ and that the lemma holds whenever a phylogenetic network on X displays $\mathcal{T}$ and has fewer reticulations than k. Let $\mathcal{S}$ be an embedding of $\mathcal{T}$ in $\mathcal{N}$, and let v be a reticulation in $\mathcal{N}$ such that no other reticulation in $\mathcal{N}$ is an ancestor of v. Since $\mathcal{N}$ is acyclic, such a reticulation exists. Let u and u' be the parents of v in $\mathcal{N}$. Note that u and u' are both tree vertices but may not be distinct. As $\mathcal{T}$ contains no vertex of in-degree two, at least one of (u, v) and (u', v) is not in the edge set $E(\mathcal{S})$ of $\mathcal{S}$. Without loss of generality, we may assume that $(u, v) \notin E(\mathcal{S})$.

Let $\mathcal{N}'$ be the phylogenetic network on X obtained from $\mathcal{N}$ by a single application of (II) that deletes (u, v) and then contracts u and v. As $\mathcal{N}$ displays $\mathcal{T}$ and $(u, v) \notin E(\mathcal{S})$, it follows that $\mathcal{N}'$ displays $\mathcal{T}$. Thus, as $\mathcal{N}'$ has $k - 1$ reticulations, we have by induction that

$$d_{\mathrm{SNPR}}(\mathcal{N}', \mathcal{T}) \leq k - 1.$$

In turn, this implies that

$$d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{T}) \leq k$$

as required. □

Proposition 7.5. *Let $\mathcal{N}$ and $\mathcal{N}'$ be two rooted binary phylogenetic networks on X. Then*

$$d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{N}') \leq \min\{d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') : \mathcal{N} \text{ displays } \mathcal{T},\ \mathcal{N}' \text{ displays } \mathcal{T}'\} + k + k',$$

where k and k' are the number of reticulations in $\mathcal{N}$ and $\mathcal{N}'$, respectively.

Proof. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees displayed by $\mathcal{N}$ and $\mathcal{N}'$, respectively, such that, amongst all such pairs of trees, the rSPR-distance is minimised. By Lemma 7.4, $d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{T}) = k$ and $d_{\mathrm{SNPR}}(\mathcal{N}',\mathcal{T}') = k'$. Therefore, by Proposition 3.1,

$$\begin{aligned} d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{N}') &\leq d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{T}) + d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') + d_{\mathrm{SNPR}}(\mathcal{T}',\mathcal{N}') \\ &= k + d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') + k', \end{aligned}$$

thereby completing the proof of the proposition. □

The next corollary is an immediate consequence of Proposition 7.5.

Corollary 7.6. *Let $\mathcal{N}$ and $\mathcal{N}'$ be two phylogenetic networks on X. If there is a rooted binary phylogenetic X-tree displayed by both $\mathcal{N}$ and $\mathcal{N}'$, then*

$$d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{N}') \leq k + k',$$

where k and k' are the number of reticulations in $\mathcal{N}$ and $\mathcal{N}'$, respectively.

We end this section by establishing one further result relating the SNPR-distance for phylogenetic networks to the rSPR-distance for rooted binary phylogenetic trees.

Proposition 7.7. *Let $\mathcal{N}$ and $\mathcal{N}'$ be two phylogenetic networks on X such that $d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{N}') = k$. Suppose that $\mathcal{T}$ is a rooted binary phylogenetic X-tree displayed by $\mathcal{N}$. Then there is a rooted binary phylogenetic X-tree $\mathcal{T}'$ displayed by $\mathcal{N}'$ such that $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') \leq k$.*

Proof. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree displayed by $\mathcal{N}$. The proof is by induction on k. If $k = 0$, then the proposition trivially holds as $\mathcal{N} \cong \mathcal{N}'$ and so $\mathcal{N}'$ also displays $\mathcal{T}$. Suppose that $k = 1$. If $\mathcal{T}$ is displayed by $\mathcal{N}'$, then choose $\mathcal{T}'$ to be $\mathcal{T}$, and we have $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = 0$. Thus we may assume that $\mathcal{T}$ is not displayed by $\mathcal{N}'$, in which case, $d_{\mathrm{SNPR}}(\mathcal{N},\mathcal{N}') = 1$ and $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by applying either (I) or (II). If (III) had been applied, then $\mathcal{T}$ is displayed by $\mathcal{N}'$; a contradiction.

Say $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by applying (I). Let $e = (u, v)$ denote the edge in $\mathcal{N}$ that is deleted in performing this operation, and let f denote the edge that results from contracting u. Let $\mathcal{S}$ be an embedding of $\mathcal{T}$ in $\mathcal{N}$. Since $\mathcal{N}'$ does not display $\mathcal{T}$, it follows that $e \in E(\mathcal{S})$. Let $\mathcal{S}'$ denote the directed subgraph of $\mathcal{N}'$ with vertex set $V(\mathcal{S}) - \{u\}$, and whose edge set is $E(\mathcal{S}) - \{e\}$ if $E(\mathcal{S})$ does not include each of the edges incident with u in $\mathcal{N}$ and $(E(\mathcal{S}) - \{e\}) \cup \{f\}$ otherwise. Let P' be a directed path in $\mathcal{N}'$ starting at a vertex in $\mathcal{S}'$, ending at v, and containing no other vertices in $\mathcal{S}'$. Note that such a path exists. It is now easily seen that the minimal directed subgraph of the subgraph of $\mathcal{N}'$ induced by the union of the vertex sets of $\mathcal{S}'$ and P', and the union of the edge sets of $\mathcal{S}'$ and P' contains X and the root of $\mathcal{N}'$ and is an embedding of a rooted binary phylogenetic X-tree $\mathcal{T}'$ in $\mathcal{N}'$. Furthermore, by construction, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$. A similar argument works if $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by applying (II). Hence the proposition holds for $k = 1$.

Now suppose that $k \geq 2$ and the proposition holds whenever the SNPR-distance between two phylogenetic networks on X is at most $k - 1$. Since $d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{N}') = k$, there is an SNPR-sequence

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k = \mathcal{N}'.$$

Consider $\mathcal{N}_{k-1}$. Now $d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{N}_{k-1}) = k - 1$ and $d_{\mathrm{SNPR}}(\mathcal{N}_{k-1}, \mathcal{N}') = 1$; otherwise, $d_{\mathrm{SNPR}}(\mathcal{N}, \mathcal{N}') < k$. Thus, by induction, there are rooted binary phylogenetic X-trees $\mathcal{T}''$ and $\mathcal{T}'$ displayed by $\mathcal{N}_{k-1}$ and $\mathcal{N}'$, respectively, such that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}'') \leq k - 1$ and $d_{\mathrm{rSPR}}(\mathcal{T}'', \mathcal{T}') \leq 1$. It follows that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq k$, thereby completing the proof of the proposition. □

8. Not Allowing Parallel Edges

As mentioned earlier, it is natural to restrict phylogenetic networks to have no edges in parallel. After all, edges represent lines of descent and two edges in parallel are simply representing the same lines. However, in the context of this paper, imposing this restriction would mean that either (i) we can only apply an operation to a phylogenetic network that does not result in any edges in parallel or (ii) if we apply an operation and edges in parallel result, then we have to modify the resulting directed graph so that it becomes a phylogenetic network with no parallel edges. Option (i) appears too restrictive.

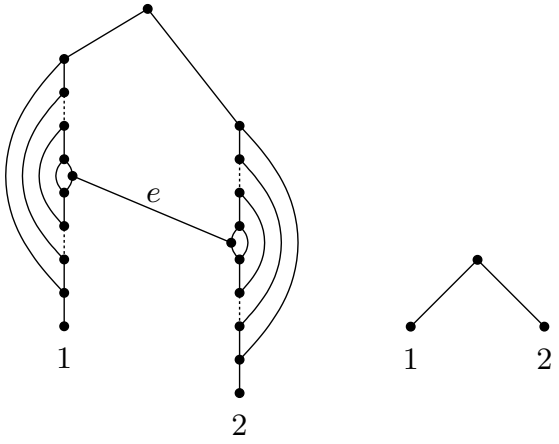

Figure 6: Two phylogenetic networks on two leaves. The right phylogenetic network is obtained from the left phylogenetic network by deleting e and, repeatedly, deleting one of two edges in parallel and contracting the resulting degree-two vertices.

Consider (ii). The canonical way to make this modification is to repeatedly do the following until there are no edges in parallel: delete one of two edges in parallel and contract the two resulting degree-two vertices. But then the deletion of a single edge has the potential to collapse much of the topology of a phylogenetic network. For example, if we delete the edge e of the phylogenetic network shown in the left of Fig. 6 and then, repeatedly, delete one of two edges in parallel and contract the resulting degree-two vertices, we obtain the phylogenetic network shown in the right of the same figure. The simple switching of a topological parent, the motivation for subnet prune and regraft, is completely lost under such a collapse. As we will shortly see, it is possible to specify exactly an operation that reverses these collapses, thereby enabling one to work with this restriction. But the reverse operation of deleting an edge would allow us to start with a phylogenetic network with no reticulations (that is, a rooted binary phylogenetic tree) and, in a single operation, obtain a phylogenetic network with an arbitrary number of reticulations.

Despite these possible sudden changes in the structure of a phylogenetic network, an operation that does not allow for parallel edges is likely to be of interest. In this section, we describe one way in which one can proceed when working under (ii).

Let $\mathcal{N}$ be a phylogenetic network on X, and let $e = (u, v)$ be an edge in $\mathcal{N}$. Consider operations (I)–(III), which we repeat here for convenience:

(I) If u is a tree vertex (and not equal to ρ), then delete e, contract u, subdivide an edge that is not a descendant of v with a new vertex u', and add the new edge (u', v).

(II) If u is a tree vertex and v is a reticulation, then delete e, and contract u and v.

(III) Subdivide e with a new vertex v', subdivide an edge f in $\mathcal{N}$ with $e \neq f$ that is not a descendant of v' with a new vertex u', and add the new edge (u', v').

Note that (III) has been slightly modified in comparison to the definition given in Section 3. This modification is so that no pair of parallel edges can be created by applying (III). Now, of the three operations, only (I) and (II) may result in a pair of parallel edges, in which case, as above, the canonical way to recover a phylogenetic network without parallel edges is to repeatedly delete one of two edges in parallel and contract the resulting degree-two vertices until there are no parallel edges. We refer to this iterative process as *unravelling parallel edges*. However, as it stands, these operations together with this process are not reversible. To resolve this issue, we need to understand how the unravelling of parallel edges can occur. The next lemma details what happens.

Lemma 8.1. *Let $\mathcal{N}$ be a phylogenetic network on X, and let $e = (u, v)$ be an edge in $\mathcal{N}$ such that u is a tree vertex. If v is a reticulation, let $\{z, z'\} = \{u, v\}$; otherwise, let $z = u$. Consider the directed graph D_z obtained from $\mathcal{N}$ by deleting e, contracting z (but not z' if it exists), and applying the iterative process of unravelling parallel edges. Then*

(i) *at each iteration i, at most one edge is deleted, in which case, two vertices, w_i and w'_i say, are contracted, where w_i is a tree vertex and w'_i is a reticulation in $\mathcal{N}$; and*

(ii) *the contracted vertices including z induce the directed path*

$$w_k, w_{k-1}, \ldots, w_1, z, w'_1, w'_2, \ldots, w'_k$$

in $\mathcal{N}$, where k is the number of iterations.

Furthermore, if z' exists and we consider the same iterative process applied to the directed graph obtained from D_z by contracting z', then (i) and (ii) hold analogously but with the induced path in D_z and not $\mathcal{N}$.

Proof. We establish (i) and (ii) for z. The proof of the second part of the lemma is similar and omitted. Let D_0 denote the directed graph obtained from $\mathcal{N}$ by deleting e and contracting z. For all $i \in \{1, 2, \ldots, k\}$, let D_i denote the directed graph at the end of iteration i. First observe that, as $\mathcal{N}$ has no parallel edges, the directed graph obtained from $\mathcal{N}$ by deleting e and contracting neither u nor v has no edges in parallel. Thus any pair of parallel edges in D_0 must include the edge resulting from the contraction of z. Therefore D_0 has at most one pair of parallel edges, in which case, denoting the end vertices of these edges by w_1 and w_1', it follows that

$$w_1,\ z,\ w_1'$$

is a directed path in $\mathcal{N}$ with w_1 a tree vertex and w_1' a reticulation.

Now consider D_1. If D_1 has no parallel edges, the process stops and the lemma is established. Suppose D_1 has a pair of parallel edges. Since the directed graph obtained from D_0 by deleting one of the two edges in parallel and contracting just w_1 has no edges in parallel, it follows that any pair of parallel edges in D_1 includes the edge resulting from the contraction of w_1 and w_1'. Therefore D_1 has at most one pair of edges in parallel, in which case, denoting the end vertices of these edges by w_2 and w_2', it follows that

$$w_2,\ w_1,\ z,\ w_1',\ w_2'$$

is a directed path in $\mathcal{N}$ with w_2 a tree vertex and w_2' a reticulation. If D_2 has no edges in parallel, the process stops. Otherwise, we can continue this argument. The process can only continue while we have an underlying cycle and so it eventually stops, and we establish the lemma. □

With Lemma 8.1 in mind, we next describe extensions of (I)–(III) that do not allow for parallel edges but, as we will show, are reversible. To this end, let D be a directed graph and suppose that w is a vertex with in-degree 1 and out-degree 1, and with incident edges (u, w) and (w, v). Let D' be the directed graph obtained from D by subdividing (u, w) with the vertices $u_1, u_2, \ldots, u_k$ so that

$$u_k,\ u_{k-1},\ \ldots,\ u_1$$

is a directed path, subdividing (w, v) with the vertices $v_1, v_2, \ldots, v_k$ so that

$$v_1,\ v_2,\ \ldots,\ v_k$$

is a directed path and, for each $i \in \{1, 2, \ldots, k\}$, adding the edge (u_i, v_i). Note that u_i has in-degree one and out-degree two whereas v_i has in-degree two and out-degree one in D'. We say that D' has been obtained from D by *adding k edges in parallel to w* or, simply, *adding edges in parallel to w* if the number of edges does not play a particular role.

Now let $\mathcal{N}$ be a phylogenetic network on X. Let $e = (u, v)$ be an edge in $\mathcal{N}$. The above-mentioned extensions of (I)–(III) are (I)*–(III)*, respectively:

(I)* If u is a tree vertex (and not equal to ρ), then delete e, contract u and unravel parallel edges, subdivide an edge that is not a descendant of v with a new vertex u', add a (possibly empty) set of new edges in parallel to u', and add the new edge (u', v).

(II)* If u is a tree vertex and v is a reticulation, then delete e, contract u and v, and unravel parallel edges.

(III)* Subdivide e with a new vertex v' and add a (possibly empty) set of new edges in parallel to v', subdivide an edge that is not a descendant of v' with a new vertex u' and add a (possibly empty) set of new edges in parallel to u', and add the new edge (u', v').

It is easily seen that applying any one of (I)*–(III)* to $\mathcal{N}$ results in a phylogenetic network on X. We say that a phylogenetic network on X has been obtained from $\mathcal{N}$ by a single SNPR *with unravelling* if it can be obtained by applying exactly one of (I)*–(III)*.

The next proposition shows that each of the operations (I)*–(III)* is indeed reversible.

Proposition 8.2. *Let $\mathcal{N}$ and $\mathcal{N}'$ be two phylogenetic networks on X. Suppose that $\mathcal{N}'$ is obtained from $\mathcal{N}$ by applying exactly one of* (I)*–(III)*. *Then, up to isomorphism, $\mathcal{N}$ can be obtained from $\mathcal{N}'$ by applying exactly one of* (I)*–(III)*.*

Proof. We prove the proposition for when $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by applying (II)*. If $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by applying either (I)* or (III)*, then similar proofs show that, up to isomorphism, $\mathcal{N}$ can be obtained from $\mathcal{N}'$ by applying (I)* or (II)*, respectively.

Let $e = (u, v)$ be an edge in $\mathcal{N}$ such that u is a tree vertex and v is a reticulation. Suppose that $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by using e in an application of (II)*. Without loss of generality, we may assume that $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by deleting e, first contracting u (but not v) and unravelling parallel edges, and then contracting v and unravelling parallel edges. Let D_u denote the directed graph obtained immediately prior to contracting v in this sequence of operations. Furthermore, including u, let U denote the set of contracted vertices in obtaining D_u from $\mathcal{N}$ and, including v, let V denote the set of contracted vertices in obtaining $\mathcal{N}'$ from D_u.

By Lemma 8.1, we can order the vertices,

$$u_k,\, u_{k-1},\, \ldots,\, u_1,\, u,\, u'_1,\, u'_2,\, \ldots,\, u'_k$$

say, in U so that the ordering induces a directed path in $\mathcal{N}$. Furthermore, by the same lemma, we can order the vertices,

$$v_l,\, v_{l-1},\, \ldots,\, v_1,\, v,\, v'_1,\, v'_2,\, \ldots,\, v'_l$$

say, in V so that the ordering induces a directed path in D_u. Let e_v denote the edge in $\mathcal{N}'$ that resulted from the contraction of v_l and v'_l, or v if contracting v resulted in no parallel edges. Then the directed graph obtained from $\mathcal{N}'$ by subdividing e_v with a new vertex v' and adding l new edges in parallel to v' is isomorphic to D_u. Now let e_u denote the edge in D_u that resulted from the contraction of u_k and u'_k, or u if contracting u resulted in no parallel edges. It is easily seen that e_u is not a descendant of v in D_u. Then the directed graph obtained from D_u by subdividing e_u with a new vertex u', adding k edges in parallel to u', and adding the new edge (u', v') is isomorphic to $\mathcal{N}$. It now follows that, up to isomorphism, $\mathcal{N}$ can be obtained from $\mathcal{N}'$ by applying a single application of (III)*. □

The definition of a SNPR-sequence using (I)*–(III)* instead of (I)–(III) is defined in the obvious way. The proof of the next proposition is similar to that of the proof of Proposition 3.2 and is omitted.

Proposition 8.3. *Let $\mathcal{N}$ and $\mathcal{N}'$ be two phylogenetic networks on X. Then there is a sequence of operations* (I)*–(III)* *connecting $\mathcal{N}$ and $\mathcal{N}'$.*

We define the SNPR *with unravelling distance* between two arbitrary phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ on X, denoted by $d_{\mathrm{SNPR}*}(\mathcal{N}, \mathcal{N}')$, to be the minimum length of a SNPR-sequence with unravelling connecting $\mathcal{N}$ and $\mathcal{N}'$. The next corollary follows immediately from Propositions 8.2 and 8.3.

Corollary 8.4. *The distance $d_{\mathrm{SNPR}*}$ is a metric on the class of all phylogenetic networks on X.*

9. Concluding Remarks

In this paper, we have presented a new rearrangement operation on rooted phylogenetic networks—called SNPR—that can transform any phylogenetic network into any other such network by a sequence of SNPR operations. The operation has a similar flavour as the rSPR operation on rooted binary phylogenetic trees in that a single SNPR moves a subnetwork across an arbitrary distance (i.e. an arbitrary number of edges) in the network, thereby switching, deleting, or inserting precisely one parent of the subnetwork. Indeed, SNPR on phylogenetic networks is a generalisation of the commonly-used rSPR operation on phylogenetic trees that is used in searching an optimal tree in practice. We have shown that several spaces of phylogenetic networks (e.g. tree child, reticulation visible, and tree based) are connected under SNPR regardless of whether or not the number of reticulations is fixed in such a space. Moreover, phylogenetic networks that display a given set of phylogenetic trees are also connected under this new operation. Related results for other classes of networks can be found in Klawitter (forthcoming). For example, for a fixed number of reticulations, the biologically relevant class of temporal networks (Moret et al., 2004) that impose several time constraints is also connected. Consequently, in reconstructing phylogenetic networks from molecular sequence data, it is possible to search through subspaces instead of the vast (infinite) space of all phylogenetic networks; in particular if one has some information on how frequent reticulation events happened for a given data set. Hence, our connectedness results are likely to play an important role in the development of new methods to analyse complex evolutionary histories that are more realistically represented by a network rather than a tree under a maximum likelihood or Bayesian-type framework. In future work, it would be interesting to extend the current SNPR framework so that it allows for a distinction between the two reticulation edges that are directed into a reticulation as it would, for example, be desirable in the context of horizontal gene transfer (Cardona et al., 2015).

Acknowledgements. The second and third author thank the New Zealand Marsden Fund for their financial support. We thank Jonathan Klawitter and an anonymous referee for their helpful comments.

Allen, B.L., Steel, M., 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics 5, 1–15.

Bordewich, M., Semple, C., 2004. On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combinatorics 8, 409–423.

Bordewich, M., Semple, C., 2016. Reticulation-visible networks. Advances in Applied Mathematics 78, 114–141.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10:e1003537.

Bryant, D., 2004. The splits in the neighborhood of a tree. Annals of Combinatorics 8, 1–11.

Cardona, G., Llabrés, M., Rosselló, F., Valiente, G., 2009a. Metrics for phylogenetic networks I: generalizations of the Robinson-Foulds metric. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 46–61.

Cardona, G., Pons, J. C., and Rosselló, F., 2015. A reconstruction problem for a class of phylogenetic networks with lateral gene transfer. Algorithms for Molecular Biology 10, 28.

Cardona, G., Rosselló, F., Valiente, G., 2009b. Comparison of tree-child phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 552–569.

Francis, A.R., Steel, M., 2015. Which phylogenetic networks are merely trees with additional arcs? Systematic Biology 64, 768–777.

Fischer, M., van Iersel, L., Kelk, S., Scornavacca, C., 2015. On computing the maximum parsimony score of a phylogenetic network. SIAM Journal on Discrete Mathematics 29, 559–585.

Gordon, K., Ford, E., St. John, K., 2013. Hamiltonian walks of phylogenetic treespaces. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10, 1076–1079.

Gusfield, D., 2014. Recombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks. MIT Press.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59, 307–321.

Huber, K.T., Linz, S., Moulton, V., Wu, T., 2016a. Spaces of phylogenetic networks from generalized nearest-neighbor interchange operations. Journal of Mathematical Biology 72, 699–725.

Huber, K.T., Moulton, V., Wu, T., 2016b. Transforming phylogenetic networks: moving beyond tree space. Journal of Theoretical Biology 404, 30–39.

Huson, D.H., Rupp, R., Scornavacca, C., 2010. Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press.

Jin, G., Nakhleh, L., Snir, S., Tuller, T., 2006a. Maximum likelihood of phylogenetic networks. Bioinformatics 22, 2604–2611.

Jin, G., Nakhleh, L., Snir, S., Tuller, T., 2006b. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Molecular Biology and Evolution, 24, 324–337.

Jin, G., Nakhleh, L., Snir, S., Tuller, T., 2009. Parsimony score of phylogenetic networks: hardness results and a linear-time heuristic. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 495–505.

Klawitter, J. Spaces of phylogenetic networks. PhD thesis, University of Auckland, *in preparation.*

Moret, B. M. E., Nakhleh, L., Warnow, T., Linder, C. R., Tholse, A., Padolina, A., Sun, J., and Timme R., 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1, 13–23.

Nakhleh, L., 2010. A metric on the space of reduced phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7, 218–222.

Owen, M., Provan, J.S., 2011. A fast algorithm for computing geodesic distances in tree space. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8, 2–13.

Robinson, D.F., 1971. Comparison of labeled trees with valency three. Journal of Combinatorial Theory, Series B 11, 105–119.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Sanderson, M.J., McMahon, M.M., Steel, M., 2011. Terraces in phylogenetic tree space. Science 333, 448–450

Semple, C., 2016. Phylogenetic networks with every embedded phylogenetic tree a base tree. Bulletin of Mathematical Biology 78, 132–137.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference in molecular systematics. Sinauer Associates.

Whidden, C., Matsen, F.A. IV, 2015. Quantifying MCMC exploration of phylogenetic tree space. Systematic Biology 64, 472–491.

Yu, Y., Barnett, R.M, Nakhleh, L., 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Systematic Biology 62, 738–751.

Yu, Y., Dong, J., Liu, K.J., Nakhleh, L., 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences of the United States of America 111, 16448–16453.